SpecInfer: Accelerating Large Language Model Serving with Tree-based Speculative Inference and Verification

Year: 2023 | Citations: 238 | Authors: Xupeng Miao, Gabriele Oliaro, Zhihao Zhang

---

Abstract

This paper introduces SpecInfer, a system that accelerates generative large language model (LLM) serving with tree-based speculative inference and verification. The key idea behind SpecInfer is leveraging small speculative models to predict the LLM's outputs; the predictions are organized as a token tree, whose nodes each represent a candidate token sequence. The correctness of all candidate token sequences represented by a token tree is verified against the LLM in parallel using a novel tree-based parallel decoding mechanism. SpecInfer uses an LLM as a token tree verifier instead of an incremental decoder, which significantly reduces the end-to-end latency and computational requirement for serving generative LLMs while provably preserving model quality. Our evaluation shows that SpecInfer outperforms existing LLM serving systems by 1.5-2.8× for distributed LLM inference and by 2.6-3.5× for offloading-based LLM inference, while preserving the same generative performance. SpecInfer is publicly available at https://github.com/flexflow/FlexFlow/